**Data Science Interview Questions - Statistics**

*Source: Interview Preparation*

Statistical Concepts for Data Science Interviews:  
   
 1. What is a p-value and how do you interpret it?  
 A p-value is the probability of observing data as extreme as the test statistic under the null hypothesis.   
 A low p-value (< 0.05) suggests rejecting the null hypothesis.  
   
 2. Explain Type I and Type II errors.  
 Type I error: rejecting a true null hypothesis (false positive).  
 Type II error: failing to reject a false null hypothesis (false negative).  
   
 3. What is the difference between correlation and causation?  
 Correlation measures the relationship between variables. Causation implies one variable directly affects another.   
 Correlation doesn't imply causation - correlation can be spurious or due to confounding variables.  
   
 4. Explain the central limit theorem.  
 The CLT states that the sampling distribution of the mean approaches a normal distribution   
 as sample size increases, regardless of the original distribution.  
   
 5. What is the difference between parametric and non-parametric tests?  
 Parametric tests assume specific distributions (e.g., normal) and parameters.   
 Non-parametric tests make fewer assumptions about the data distribution.  
   
 6. How do you handle missing data?  
 Methods include: deletion, imputation (mean/median/mode), advanced imputation (k-NN, regression),   
 multiple imputation, domain-specific approaches.  
   
 7. What is multicollinearity and how do you detect it?  
 Multicollinearity occurs when predictor variables are highly correlated.   
 Detection: correlation matrix, VIF (Variance Inflation Factor), condition number.  
   
 8. Explain the concept of statistical power.  
 Statistical power is the probability of correctly rejecting a false null hypothesis.   
 It depends on sample size, effect size, and significance level.  
   
 9. What is the difference between parametric and non-parametric correlation?  
 Parametric correlation (Pearson) assumes linear relationship and normal distribution.   
 Non-parametric correlation (Spearman, Kendall) doesn't require these assumptions.  
   
 10. How do you determine sample size for a study?  
 Consider: desired power, effect size, significance level, population variance,   
 practical constraints, and study design.